

Learning Sparse Low-Threshold Linear Classifiers

Sivan Sabato

SIVAN_SABATO@MICROSOFT.COM

*Microsoft Research New England
1 Memorial Drive
Cambridge, MA*

Shai Shalev-Shwartz

SHAIS@CS.HUJI.AC.IL

*Benin school of Computer Science and Engineering
The Hebrew University
Givat Ram, Jerusalem 91904, Israel*

Nathan Srebro

NATI@TTI.EDU

*Toyota Technological Institute at Chicago
6045 S. Kenwood Ave.
Chicago, IL 60637, USA*

Daniel Hsu

DAHSU@MICROSOFT.COM

*Microsoft Research New England
1 Memorial Drive
Cambridge, MA*

Tong Zhang

TZHANG@STAT.RUTGERS.EDU

*Department of Statistics
Rutgers University
Piscataway, NJ, 08854*

Abstract

We consider the problem of learning a non-negative linear classifier with 1-norm of at most k and a fixed threshold, under the hinge-loss. This problem generalizes the problem of learning a k -monotone disjunction. We prove that we can learn efficiently in this setting, at a rate which is linear in both k and the size of the threshold, and that this is the best possible rate. We provide an efficient online learning algorithm that achieves the optimal rate, and show that in the batch case, empirical risk minimization achieves this rate as well. The rates we show are tighter than the uniform convergence rate, which grows with k^2 .

Keywords: linear classifiers, monotone disjunctions, online learning, empirical risk minimization, uniform convergence

1. Introduction

We consider the problem of learning non-negative, low- ℓ_1 -norm linear classifiers *with a fixed (or bounded) threshold*. That is, we consider hypothesis classes over instances $x \in [0, 1]^d$ of the following form:

$$\mathcal{H}_{k,\theta} = \left\{ x \mapsto \langle w, x \rangle - \theta \mid w \in \mathbb{R}_+^d, \|w\|_1 \leq k \right\}, \quad (1)$$

where we associate each (real valued) linear predictor in $\mathcal{H}_{k,\theta}$ with a binary classifier:¹

$$x \mapsto \text{sign}(\langle w, x \rangle - \theta) = \begin{cases} 1 & \text{if } \langle w, x \rangle > \theta \\ -1 & \text{if } \langle w, x \rangle < \theta \end{cases}. \quad (2)$$

Note that the hypothesis class is specified by both the ℓ_1 -norm constraint k and the fixed threshold θ . In fact, the main challenge here is to understand how the complexity of learning $\mathcal{H}_{k,\theta}$ changes with θ .

The classes $\mathcal{H}_{k,\theta}$ can be seen as a generalization and extension of the class of k -monotone-disjunctions and r -of- k -formulas. Considering binary instances $x \in \{0, 1\}^d$, the class of k -monotone-disjunctions corresponds to linear classifiers with binary weights, $w \in \{0, 1\}^d$, with $\|w\|_1 \leq k$ and a fixed threshold of $\theta = \frac{1}{2}$. That is, a restriction of $\mathcal{H}_{k,\frac{1}{2}}$ to integer weights and integer instances. More generally, the class of r -of- k formulas (i.e. formulas which are true if at least r of a specified k variables are true) corresponds to a similar restriction, but with a threshold of $\theta = r - \frac{1}{2}$.

Studying k -disjunctions and r -of- k formulas, Littlestone (1988) presented the efficient Winnow online learning rule, which entertains an online mistake bound (in the separable case) of $O(k \log d)$ for k -disjunctions and $O(rk \log d)$ for r -of- k -formulas. In fact, in his analysis, Littlestone considered also the more general case of real-valued weights, corresponding to the class $\mathcal{H}_{k,\theta}$, though still only over binary instances $x \in \{0, 1\}^d$ and only for separable data, and showed that Winnow enjoys a mistake bound of $O(\theta k \log d)$ in this case as well. By applying a standard Online-to-Batch conversion (see e.g. Shalev-Shwartz, 2012), one can also achieve a sample complexity upper bound of $O(\theta k \log(d)/\epsilon)$ for batch supervised learning of this class in the separable case.

In this paper, we consider the more general case, where the instances x can also be fractional, i.e. where $x \in [0, 1]^d$. More importantly, we consider also the agnostic, non-separable, case. In order to move on to the fractional and agnostic analysis, we must clarify the loss function we will use, and the related issue of separation with a margin.

When the instances x and weight vectors w are integer-valued, we have that $\langle w, x \rangle$ is always integer. Therefore, if positive and negative instances are at all separated by some predictor w (i.e. $\text{sign}(\langle w, x \rangle - \theta) = y$ where $y \in \{\pm 1\}$ denotes the target label), they are necessarily separated by a margin of half. That is, setting $\theta = r - \frac{1}{2}$ for an integer r , we have $y(\langle w, x \rangle - \theta) \geq \frac{1}{2}$. Moving to fractional instances and weight vectors, we need to require such a margin explicitly. And if considering the agnostic case, we must account not only for mis-classified points, but also for margin violations. As is standard both in online learning (e.g. the agnostic Perceptron guarantee in Gentile 2003) and in statistical learning using convex optimization (e.g. support vector machines), we will rely on the hinge loss at margin half,² which is equal to: $2 \cdot [\frac{1}{2} - yh(x)]_+$. The hinge loss is a convex upper bound to the zero-one loss (that is, the misclassification rate) and so obtaining learning guarantees for it translates to guarantees on the misclassification error rate.

Phrasing the problem as hinge-loss minimization over the hypothesis class $\mathcal{H}_{k,\theta}$, we can use Online Exponentiated Gradient (EG) (Kivinen and Warmuth, 1994) or Online Mirror Descent (MD) (e.g. Shalev-Shwartz, 2007; Srebro et al., 2011), which rely only on the ℓ_1 -bound and hold for any threshold. In the statistical setting, we can use Empirical Risk Minimization (ERM), in this case minimizing the empirical hinge loss, and rely on uniform concentration for bounded ℓ_1 predictors (Schapire et al., 1997; Zhang, 2002; Kakade et al., 2009), again regardless of the threshold.

1. The value of the mapping when $\langle w, x \rangle = \theta$ can be arbitrary, as our results and our analysis do not depend on it.
2. Measuring the hinge loss at a margin of half rather than a margin of one is an arbitrary choice, which corresponds to a scaling by a factor of two, which fits better with the integer case discussed above.

However, these approach yield mistake bounds or sample complexities that scale quadratically with the ℓ_1 norm, that is with k^2 rather than with θk . Since the relevant range of thresholds is $0 \leq \theta \leq k$, a scaling of θk is always better than k^2 . When θ is large, that is, roughly $k/2$, the Winnow bound agrees with the EG and MD bounds. But when we consider classification with a small threshold (for instance, $\theta = \frac{1}{2}$ in the case of disjunctions, the Winnow analysis clarifies that this is a much simpler class, with a resulting smaller mistake bound and sample complexity, scaling with k rather than with k^2 . This distinction is lost in the EG and MD analyses, and in the ERM guarantee based on uniform convergence arguments, and for small thresholds, where $\theta = O(1)$, the difference between these analyses and the Winnow guarantee is a factor of k .

Our starting point and our main motivation for this paper is to understand this gap between the EG, MD and uniform concentration analyses and the Winnow analysis. Is this gap an artifact of the integer domain or the separability assumption? Or can we obtain guarantees that scale as θk rather than k^2 also in the non-integer non-separable case? In the statistical setting, must we use an online algorithm (such as Winnow) and an online-to-batch conversion in order to ensure a sample complexity that scales with θk , or can we obtain the same sample complexity also with ERM? Is it possible to establish uniform convergence guarantees with a dependence on θk rather than k^2 , or do the learning guarantees here arise from a more delicate argument?

The gap between the Winnow analysis and the more general ℓ_1 -norm-based analyses is particularly disturbing since we know that, in a sense, online mirror descent always provides the best possible rates in the online setting (Srebro et al., 2011), and uniform concentration based guarantees provide the best possible rates for supervised learning in the PAC model (Alon et al., 1993).

Answering the above questions, our main contributions are:

- We provide a variant of online Exponentiated Gradient, for which we establish a regret bound of $O(\sqrt{\theta k \log(d)T})$ for $\mathcal{H}_{k,\theta}$, improving on the $O(\sqrt{k^2 \log(d)T})$ regret guarantee ensured by the standard EG analysis. We do so using a more refined analysis based on local norms (Section 3). Using a standard online-to-batch conversion, this yields a sample complexity of $O(\theta k \log(d)/\epsilon^2)$ in the statistical setting.
- In the statistical agnostic PAC setting, we show that the rate of uniform convergence of the empirical hinge loss of predictors in $\mathcal{H}_{k,\theta}$ is indeed $\Omega(\sqrt{k^2/m})$ where m is the sample size, corresponding to a sample complexity of $\Omega(k^2/\epsilon^2)$, even when θ is small (Section 5). Nevertheless, we establish a learning guarantee for empirical risk minimization which matches the online-to-batch guarantee above (up to logarithmic factors), and ensures a sample complexity of $\tilde{O}(\theta k \log(d)/\epsilon^2)$ also when using ERM. This is obtained by a more delicate local analysis, focusing on predictors which might be chosen as empirical risk minimizers, rather than a uniform analysis over the entire class $\mathcal{H}_{k,\theta}$ (Section 4).
- We also establish a matching lower bound (up to logarithmic factors) of $\Omega(\theta k/\epsilon^2)$ on the required sample complexity for learning $\mathcal{H}_{k,\theta}$ in the statistical setting. This shows that our ERM analysis is tight (up to logarithmic factors), and that, furthermore, the regret guarantee we obtain in the online setting is likewise tight up to logarithmic factors.

1.1 Related Prior Work

We discussed Littlestone’s work on Winnow at length above. In our notation, Littlestone (1988) established a mistake bound (that is, a regret guarantee in the separable case, where there exists a

predictor with zero hinge loss) of $O(k\theta \log(d))$ for $\mathcal{H}_{k,\theta}$, when the instances are integer $x \in \{0, 1\}^d$. Littlestone also established a lower bound of $k \log(d/k)$ on the VC-dimension of k -monotone-disjunctions, corresponding to the case $\theta = \frac{1}{2}$, thus implying a $\Omega(k \log(d/k)/\epsilon^2)$ lower bound on learning $\mathcal{H}_{k,\frac{1}{2}}$. However, the question of obtaining a lower bound for other values of the threshold θ was left open by Littlestone.

In the agnostic case, Auer and Warmuth (1998) studied the discrete problem of k -monotone disjunctions, corresponding to $\mathcal{H}_{k,\frac{1}{2}}$ with integer instances $x \in \{0, 1\}^d$ and integer weights $w \in \{0, 1\}^d$, under the *attribute loss*, defined as the number of variables in the assignment that need to be flipped in order to make the predicted label correct. They provide an online algorithm with an expected mistake bound of $A^* + 2\sqrt{A^*k \ln(d/k)} + O(k \ln(d/k))$, where A^* is the best possible attribute loss for the given online sequence. An online-to-batch conversion thus achieves here a zero-one loss which converges to the optimal attribute loss on this problem at the rate of $O(k \ln(d/k)/\epsilon^2)$. Since the attribute loss is upper bounded by the hinge loss, this result holds also when replacing A^* with the optimal hinge-loss for the given sequence. This establishes an agnostic guarantee of the desired form, for a threshold of $\theta = \frac{1}{2}$, and when both the instances and weight vectors are integer. We are not aware of work on $\mathcal{H}_{k,\theta}$ in the agnostic case for $\theta > \frac{1}{2}$ or when the instances x or the weights w are fractional.

2. Notations and definitions

For a real number q , we denote its positive part by $[q]_+ := \max\{0, q\}$. We denote universal positive constants by C . The value of C may be different between statements or even between lines of the same expression. We denote by \mathbb{R}_+^d the non-negative orthant in \mathbb{R}^d .

We will slightly overload notation and use $w \in \mathcal{H}_{k,\theta}$ to denote both the vector $w \in \mathbb{R}_+^d$ and the linear predictor $x \mapsto \langle w, x \rangle - \theta$ associated with it, where θ is implied.

For convenience we will work with *half* the hinge loss at margin half, and denote this loss, for a predictor $w \in \mathcal{H}_{k,\theta}$, for $\theta \in [0, k]$, by

$$\ell_\theta(x, y, w) := \left[\frac{1}{2} - y(\langle w, x \rangle - \theta) \right]_+.$$

The subscript θ will sometimes be omitted when it is clear from context.

Echoing the half-integer thresholds for k -monotone-disjunctions, r -of- k formulas, and the discrete case more generally, we will denote $r = \theta + \frac{1}{2}$, so that $\theta = r - \frac{1}{2}$. In the discrete case r is integer, but in this paper $\frac{1}{2} \leq r \leq k - \frac{1}{2}$ can also be fractional. We will also sometimes refer to $r' = \frac{1}{2} - \theta$. Note that r' can be negative.

In the statistical setting, we refer to some fixed and unknown distribution D over instance-label pairs (x, y) , where we assume access to a sample (training set) drawn i.i.d. from D , and the objective is to minimize the expected loss:

$$\ell_\theta(w, D) = \mathbb{E}_{x,y \sim D}[\ell_\theta(x, y, w)]. \quad (3)$$

When the distribution D is clear from context, we simply write $\ell_\theta(w)$, and we might also omit the subscript θ . For a set of predictors (hypothesis class) H , we denote $\ell_\theta^*(H, D) := \min_{w \in H} \ell_\theta(w, D)$. For a sample $S \in ([0, 1]^d \times \{\pm 1\})^*$, we use the notation

$$\hat{\mathbb{E}}_S[f(Z)] = \frac{1}{|S|} \sum_{i=1}^{|S|} f(S_i) \quad (4)$$

and again sometimes drop the subscript S when it is clear from context.

2.1 Rademacher complexity

The empirical Rademacher complexity of the Winnow loss for a class $W \subseteq \mathbb{R}^d$ with respect to a sample $S = ((x_1, y_1), \dots, (x_m, y_m)) \in ([0, 1]^d \times \{\pm 1\})^m$ is

$$\mathcal{R}(W, S) := \frac{2}{m} \mathbb{E} \left[\sup_{w \in W} \left| \sum_{i=1}^m \epsilon_i \ell(x_i, y_i, w) \right| \right] \quad (5)$$

where the expectation is over $\epsilon_1, \dots, \epsilon_m$ which are independent random variables drawn uniformly from $\{\pm 1\}$. The average Rademacher complexity of the Winnow loss for a class $W \subseteq \mathbb{R}^d$ with respect to a distribution D over $[0, 1]^d \times \{\pm 1\}$ is denoted by

$$\mathcal{R}_m(W, D) := \mathbb{E}_{S \sim D^m} [\mathcal{R}(W, S)] \quad (6)$$

We also define the average Rademacher complexity of W with respect to the *linear loss* by

$$\mathcal{R}_m^L(W, D) := \frac{2}{m} \mathbb{E} \left[\sup_{w \in W} \left| \sum_{i=1}^m \epsilon_i Y_i \langle w, X_i \rangle \right| \right] \quad (7)$$

where the expectation is over $\epsilon_1, \dots, \epsilon_m$ as above and $((X_1, Y_1), \dots, (X_m, Y_m)) \sim D^m$.

2.2 Probability tools

We use the following form of Bernstein's inequality: For a random variable $X \in \{0, 1\}$, with probability at least $1 - \delta$ over n i.i.d. draws of X ,

$$\hat{\mathbb{E}}[X] - \mathbb{E}[X] \leq 2 \sqrt{\frac{\ln(1/\delta)}{n} \cdot \max \left(\mathbb{E}[X], \frac{\ln(1/\delta)}{n} \right)}. \quad (8)$$

The same holds for $\mathbb{E}[X] - \hat{\mathbb{E}}[X]$.

We further use the following lemma, which bounds the ratio between the empirical fraction of positive or negative labels and their true probabilities. We will apply this lemma make sure that enough negative and positive labels can be found in a random sample.

Lemma 1 *Let B be a binomial random variable, $B \sim \text{Binomial}(m, p)$. if*

$$p \geq \frac{16 \ln(1/\delta)}{m} \quad (9)$$

then with probability of at least $1 - \delta$, $B \geq mp/2$.

Proof Denote $\hat{p} = B/m$. From Bernstein's inequality (Eq. (8)), with probability of at least $1 - \delta$:

$$\hat{p} \geq p - 2 \sqrt{\frac{\ln(1/\delta)}{m} \max(p, \frac{\ln(1/\delta)}{m})}$$

Under Eq. (9), we have that $\max(p, \frac{\ln(1/\delta)}{m}) = p$ and that $\frac{\ln(1/\delta)}{pm} \leq \frac{1}{16}$, which yields

$$\hat{p} \geq p - 2\sqrt{\frac{p \ln(1/\delta)}{m}} = p \left(1 - 2\sqrt{\frac{\ln(1/\delta)}{pm}} \right) \geq p \left(1 - 2\sqrt{\frac{1}{16}} \right) = \frac{p}{2}.$$

■

3. Online Algorithm

Consider the following algorithm:

**Unnormalized Exponentiated Gradient
(unnormalized-EG)**

parameters: $\eta, \lambda > 0$
input: $\mathbf{z}_1, \dots, \mathbf{z}_T \in \mathbb{R}^d$
initialize: $\mathbf{w}_1 = (\lambda, \dots, \lambda) \in \mathbb{R}^d$
update rule: $\forall i, w_{t+1}[i] = w_t[i]e^{-\eta z_t[i]}$

The following theorem provides a regret bound with local-norms for the unnormalized EG algorithm. For a proof see Shalev-Shwartz (2012), Theorem 2.23.

Theorem 2 *Assume that the unnormalized EG algorithm is run on a sequence of vectors such that for all t, i we have $\eta z_t[i] \geq -1$. Then, for all $\mathbf{u} \geq \mathbf{0}$,*

$$\sum_{t=1}^T \langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t \rangle \leq \frac{d\lambda + \sum_{i=1}^d u[i] \ln(u[i]/(e\lambda))}{\eta} + \eta \sum_{t=1}^T \sum_{i=1}^d w_t[i] z_t[i]^2.$$

Now, let us apply it to a case in which we have a sequence of convex functions f_1, \dots, f_T , and \mathbf{z}_t is the sub-gradient of f_t at \mathbf{w}_t . Additionally, set $\lambda = 1/d$ and consider \mathbf{u} s.t. $\|\mathbf{u}\|_1 \leq k$. We obtain

Theorem 3 *Assume that the unnormalized EG algorithm is run with $\lambda = 1/d$. Assume that for all t , we have $\mathbf{z}_t \in \partial f_t(\mathbf{w}_t)$, for some convex function f_t . Further assume that for all t, i we have $\eta z_t[i] \geq -1$, and that for some positive constants α, β we have that*

$$\sum_{i=1}^d w_t[i] z_t[i]^2 \leq \alpha f_t(\mathbf{w}_t) + \beta. \quad (10)$$

Then, for all $\mathbf{u} \geq \mathbf{0}$, with $\|\mathbf{u}\|_1 \leq k$ we have

$$\sum_{t=1}^T f_t(\mathbf{w}_t) \leq \frac{1}{1 - \alpha\eta} \left(\sum_{t=1}^T f_t(\mathbf{u}) + \frac{2k \ln(kd)}{\eta} + \eta\beta T \right).$$

Proof Using the convexity of f_t and the assumption that $\mathbf{z}_t \in \partial f_t(\mathbf{w}_t)$ we have that

$$\sum_{t=1}^T (f_t(\mathbf{w}_t) - f_t(\mathbf{u})) \leq \sum_{t=1}^T \langle \mathbf{w}_t - \mathbf{u}, \mathbf{z}_t \rangle.$$

Combining with Theorem 2 we obtain

$$\sum_{t=1}^T (f_t(\mathbf{w}_t) - f_t(\mathbf{u})) \leq \frac{d\lambda + \sum_{i=1}^d u[i] \ln(u[i]/(e\lambda))}{\eta} + \eta \sum_{t=1}^T \sum_{i=1}^d w_t[i] z_t[i]^2.$$

Using the assumption in Eq. (10), the definition of $\lambda = 1/d$, and the assumptions on \mathbf{u} , we obtain

$$\sum_{t=1}^T (f_t(\mathbf{w}_t) - f_t(\mathbf{u})) \leq \frac{2k \ln(kd)}{\eta} + \eta\beta T + \eta\alpha \sum_{t=1}^T f_t(\mathbf{w}_t).$$

Rearranging the above we conclude our proof. ■

We can now show the desired regret bound for our algorithm.

Corollary 4 Fix any sequence $(x_1, y_1), (x_2, y_2), \dots, (x_T, y_T) \in [0, 1]^d \times \{\pm 1\}$ and assume $T \geq 8k \ln(kd)/r$. Suppose the unnormalized EG algorithm listed in Section 3 is run using $\eta := \sqrt{\frac{2k \ln(kd)}{rT}}$, $\lambda := 1/d$, and any $\mathbf{z}_t \in \partial_w \ell(x_t, y_t, w_t)$ for all t . Fix any $u \in \mathbb{R}_+^d$, and define $L_{\text{UEG}} := \frac{1}{T} \sum_{t=1}^T \ell(x_t, y_t, w_t)$ and $L(u) := \frac{1}{T} \sum_{t=1}^T \ell(x_t, y_t, u)$. Then

$$L_{\text{UEG}} \leq L(u) + \sqrt{L(u)^2 \cdot \frac{8k \ln(kd)}{rT}} + \sqrt{\frac{8rk \ln(kd)}{T}} + \frac{8k \ln(kd)}{T}.$$

Proof Every sub-gradient $\mathbf{z}_t \in \partial_w \ell(x_t, y_t, w_t)$ is of the form $\mathbf{z}_t = a_t \mathbf{x}_t$ for some $a_t \in \{-1, 0, +1\}$. Since $0 \leq x_t[i] \leq 1$ and $w_t[i] \geq 0$ for all i , it follows that $\sum_{i=1}^d w_t[i] z_t[i]^2 = |a_t| \sum_{i=1}^d w_t[i] x_t[i]^2 \leq |a_t| \langle w_t, x_t \rangle$. Now consider three disjoint cases.

- Case 1: $\langle w_t, x_t \rangle \leq r$. Then $\sum_{i=1}^d w_t[i] z_t[i]^2 \leq \langle w_t, x_t \rangle \leq r$.
- Case 2: $\langle w_t, x_t \rangle > r$ and $y = 1$. Then $a_t = 0$ and $\sum_{i=1}^d w_t[i] z_t[i]^2 = 0$.
- Case 3: $\langle w_t, x_t \rangle > r$ and $y = -1$. Then $\sum_{i=1}^d w_t[i] z_t[i]^2 \leq \langle w_t, x_t \rangle \leq [r' + \langle w_t, x_t \rangle]_+ - r' \leq [r' + \langle w_t, x_t \rangle]_+ + r$.

In all three cases, the final upper bound on $\sum_{i=1}^d w_t[i] z_t[i]^2$ is at most $\ell(x_t, y_t, w_t) + r$. Therefore, Eq. (10) from Theorem 3 is satisfied with $f_t(w) := \ell(x_t, y_t, w)$, $\alpha := 1$, and $\beta := r$. The claim now follows from Theorem 3 with this choice of f_t and the given settings of η , λ , and \mathbf{z}_t (using the inequality $1/(1-x) \leq 1+2x$ for $x \in [0, 1/2]$). ■

4. ERM Upper bound

We now proceed to the batch setting. We wish to show an upper bound on $\ell(\hat{w}) - \ell(w^*)$, where $w^* \in \operatorname{argmin}_{w \in W_k} \mathbb{E}[\ell(X, Y, w)]$, and $\hat{w} \in \operatorname{argmin}_{w \in W_k} \frac{1}{m} \sum_{i \in [m]} \ell(x_i, y_i, w)$ is an ERM. We will prove the following theorem:

Theorem 5 *For $k \geq r \geq 0$, with probability $1 - \delta$*

$$\ell(\hat{w}) \leq \ell(w^*) + \sqrt{\frac{O(rk(\ln(kd) \ln^3(m) + \ln(1/\delta)))}{m}}. \quad (11)$$

Our proof strategy will be to consider the loss on negative examples and the loss on positive examples separately. Denote

$$\begin{aligned} \ell_-(w, D) &= \mathbb{E}_{(X, Y) \sim D}[\ell(X, Y, w) \mid Y = -1], \text{ and} \\ \ell_+(w, D) &= \mathbb{E}_{(X, Y) \sim D}[\ell(X, Y, w) \mid Y = +1]. \end{aligned}$$

For a given sample $((X_1, Y_1), \dots, (X_m, Y_m))$, Denote $\hat{\ell}_-(w) = \hat{\mathbb{E}}[\ell(X, Y, w) \mid Y = -1]$ and similarly for $\hat{\ell}_+(w)$. As we show in Section 5.2, uniform convergence for negative examples is too slow if we consider any $w \in W_k$. However, we will show that the rate is fast enough for any w that might be returned by an algorithm that minimizes the loss on a sample drawn from D . For positive labels, we will show that with high probability over the draw of an i.i.d. sample from D , the true loss of any $w \in W_k$ on examples with positive labels is close to the empirical loss of that w on positive examples. We will then combine the two results while taking into account the balance between positive and negative labels in D .

4.1 Convergence on Negative labels

We now commence our proof for the convergence rate of ERM for the Winnow loss. As shown in Theorem 21, the empirical Winnow loss for negative examples does not converge fast enough to the true loss on negative examples for all $w \in W_k$. Luckily, not all $w \in W_k$ might be returned by an algorithm that minimizes the Winnow loss. We now show that with high probability the output of the ERM algorithm belongs to a more restricted class than W_k . Fix a sample $((x_1, y_1), \dots, (x_m, y_m))$, and let

$$\hat{w} \in \operatorname{argmin}_{w \in W_k} \frac{1}{m} \sum_{i \in [m]} \ell(x_i, y_i, w).$$

We first show a sample-dependent restriction on \hat{w} .

For a given distribution D , denote $p_+ = \mathbb{E}_{(X, Y) \sim D}[Y = +1]$ and $\hat{p}_+ = \hat{\mathbb{E}}[Y = +1]$, and similarly for p_- and \hat{p}_- .

Lemma 6

$$\hat{\mathbb{E}}[\langle \hat{w}, X \rangle \mid Y = -1] \leq \frac{r}{\hat{p}_-}.$$

Proof Let $m_+ = |\{i \mid y_i = +1\}|$, and $m_- = |\{i \mid y_i = -1\}|$. By the definition of the hinge function and the fact that $\langle x_i, \hat{w} \rangle \geq 0$ for all i we have that

$$\begin{aligned} m_+ r' + \sum_{y_i=-1} \langle x_i, \hat{w} \rangle &\leq \sum_{y_i=-1} (r' + \langle x_i, \hat{w} \rangle) \\ &\leq \sum_{y_i=+1} [r - \langle x_i, \hat{w} \rangle]_+ + \sum_{y_i=-1} [r' + \langle x_i, \hat{w} \rangle]_+ \\ &= \sum_{i \in [m]} \ell(x_i, y_i, \hat{w}). \end{aligned}$$

By the optimality of \hat{w} ,

$$\sum_{i \in [m]} \ell(x_i, y_i, \hat{w}) \leq \sum_{i \in [m]} \ell(x_i, y_i, \mathbf{0}) = m_- r + m_+ [r']_+.$$

Therefore

$$\sum_{y_i=-1} \langle x_i, \hat{w} \rangle \leq m_- r + m_+ ([r']_+ - r') = m_- r + m_+ [-r']_+ \leq (m_- + m_+) r = mr.$$

Dividing both sides by m_- we conclude our proof. ■

The next lemma will allow us to conclude from Lemma 6 that \hat{w} is in a restricted class with high probability over the samples.

Lemma 7 *For any distribution over $[0, 1]^d$, with probability $1 - \delta$ over samples of size n , for any $w \in W_k$*

$$\mathbb{E}[\langle w, X \rangle] \leq 2\hat{\mathbb{E}}[\langle w, X \rangle] + \frac{16k \ln(\frac{d}{\delta})}{n}.$$

Proof For every $j \in [d]$, denote $\alpha_j = \mathbb{E}[X[j]]$. Denote $\hat{\alpha}_j = \hat{\mathbb{E}}[X[j]]$. By Bernstein's inequality (Eq. 8), with probability $1 - \delta$,

$$\alpha_j \leq \hat{\alpha}_j + 2\sqrt{\frac{\ln(1/\delta)}{n} \cdot \max\left(\alpha_j, \frac{\ln(1/\delta)}{n}\right)} \leq \hat{\alpha}_j + \max\left(\frac{\alpha_j}{2}, \frac{8 \ln(1/\delta)}{n}\right),$$

where the last inequality can be verified by considering the cases $\alpha_j \leq \frac{16 \ln(1/\delta)}{n}$ and $\alpha_j \geq \frac{16 \ln(1/\delta)}{n}$. Applying the union bound over $j \in [d]$ we obtain that with probability of $1 - \delta$ over samples of size n , for any $w \in W_k$

$$\begin{aligned} \mathbb{E}[\langle w, X \rangle] &= \langle w, \alpha \rangle \leq \sum_{j \in [d]} w_j \left(\hat{\alpha}_j + \frac{\alpha_j}{2} + \frac{8 \ln(d/\delta)}{n} \right) \\ &\leq \hat{\mathbb{E}}[\langle w, X \rangle] + \frac{1}{2} \mathbb{E}[\langle w, X \rangle] + \frac{8 \ln(d/\delta)}{n} \cdot k. \end{aligned}$$

Thus

$$\mathbb{E}[\langle w, X \rangle] \leq 2\hat{\mathbb{E}}[\langle w, X \rangle] + \frac{16k \ln(d/\delta)}{n}.$$

■

We can now conclude a restriction on \hat{w} with high probability.

Theorem 8 *If $p_- \geq \frac{16 \ln(1/\delta)}{m}$, then with probability $1 - 2\delta$ over samples of size m ,*

$$\mathbb{E}[\langle \hat{w}, X \rangle \mid Y = -1] \leq \frac{4r}{p_-} + \frac{32k \ln(d/\delta)}{mp_-}. \quad (12)$$

Proof Lemma 7 implies that with probability of $1 - \delta$ over samples drawn from D that have n negative examples,

$$\mathbb{E}[\langle w, X \rangle \mid Y = -1] \leq 2\hat{\mathbb{E}}[\langle w, X \rangle \mid Y = -1] + \frac{16k \ln(d/\delta)}{n}.$$

Therefore, by Lemma 6

$$\begin{aligned} \mathbb{E}[\langle w, X \rangle \mid Y = -1] &\leq 2\hat{\mathbb{E}}[\langle w, X \rangle \mid Y = -1] + \frac{16k \ln(d/\delta)}{m\hat{p}_-} \\ &\leq \frac{2r}{\hat{p}_-} + \frac{16k \ln(d/\delta)}{m\hat{p}_-} \\ &\leq \frac{4r}{p_-} + \frac{32k \ln(d/\delta)}{mp_-}, \end{aligned} \quad (13)$$

where the last inequality follows from the assumption and Lemma 1. ■

This theorem shows that to bound the sample complexity of an ERM algorithm, it suffices to show convergence rates of the empirical loss for w that satisfy Eq. (12). For any $b \geq 0$ and a fixed distribution D , define

$$U_b = \{w \in \mathbb{R}_+^d \mid \|w\|_1 \leq k, \mathbb{E}_D[\langle w, X \rangle] \leq b\}.$$

Note that $U_b \subseteq W_k$, and that b can be set according to Eq. (12) so that with high probability $\hat{w} \in U_b$. We bound the rate of convergence of the empirical loss on negative examples to the true loss on negative examples for all $w \in U_b$. This is accomplished in two stages: first we bound $\mathcal{R}_m^L(U_b, D)$ for any distribution D over $[0, 1]^d \times \{\pm 1\}$, and then we conclude a similar bound on $\mathcal{R}_m(U_b, D)$ for any D that draws only negative labels.

We first prove a more general lemma that we will use to derive the desired bound.

Lemma 9 *For a fixed distribution over D over $[0, 1]^d \times \{\pm 1\}$, let $\alpha_j = \mathbb{E}_{(X,Y) \sim D}[X[j]]$, and let $\mu \in \mathbb{R}^d$ be a non-negative vector. Define*

$$U^\mu = \{w \in \mathbb{R}_+^d \mid \langle w, \mu \rangle \leq 1\}.$$

then if $dm \geq 3$,

$$\mathcal{R}_m^L(U^\mu, D) \leq \max_{j: \alpha_j > 0} \frac{1}{\mu_j} \sqrt{\frac{32 \ln(d)}{m} \cdot \max \left(\alpha_j, \frac{\ln(dm)}{m} \right)}$$

Proof Assume w.l.o.g that $\alpha_j > 0$ for all j (if this is not the case, dimensions with $\alpha_j = 0$ can be removed because this implies that $X[j] = 0$ with probability 1).

$$\frac{m}{2} R_m^L(U^\mu, S) = \mathbb{E}_\sigma \left[\sup_{w: \langle w, \mu \rangle \leq 1} \sum_{i=1}^m \sigma_i \langle w, x_i \rangle \right] \quad (14)$$

$$= \mathbb{E}_\sigma \left[\sup_{w: \langle w, \mu \rangle \leq 1} \langle w, \sum_{i=1}^m \sigma_i x_i \rangle \right] \quad (15)$$

$$= \mathbb{E}_\sigma \left[\max_{j \in [d]} \sum_{i=1}^m \sigma_i \frac{x_i[j]}{\mu[j]} \right]. \quad (16)$$

Therefore, using Massart's lemma and denoting $\hat{\alpha}_j = \frac{1}{m} \sum_{i \in [m]} x_i[j]$, we have:

$$\begin{aligned} R_m^L(U^\mu, S) &\leq \frac{\sqrt{8 \ln(d)}}{m} \cdot \max_j \frac{\sqrt{\sum_i x_i[j]^2}}{\mu[j]} \\ &\leq \frac{\sqrt{8 \ln(d)}}{m} \cdot \max_j \frac{\sqrt{\sum_i x_i[j]}}{\mu[j]} \\ &= \sqrt{\frac{8 \ln(d)}{m}} \cdot \max_j \frac{\sqrt{\hat{\alpha}_j}}{\mu[j]} \\ &= \sqrt{\frac{8 \ln(d)}{m}} \cdot \max_j \frac{\hat{\alpha}_j}{\mu[j]^2}. \end{aligned}$$

Taking expectation over S and using Jensen's inequality we obtain

$$R_m^L(U^\mu, D) = \mathbb{E}_S[R_m^L(U^\mu, S)] \leq \sqrt{\frac{8 \ln(d)}{m}} \cdot \mathbb{E}_S[\max_j \frac{\hat{\alpha}_j}{\mu[j]^2}]$$

By Bernstein's inequality, with probability $1 - \delta$ over the choice of $\{x_i\}$, for all $j \in [d]$

$$\hat{\alpha}_j \leq \alpha_j + 2 \sqrt{\frac{\ln(d/\delta)}{m}} \cdot \max \left(\alpha_j, \frac{\ln(d/\delta)}{m} \right).$$

And, in any case, $\hat{\alpha}_j \leq 1$. Therefore,

$$\mathbb{E}_S[\max_j \frac{\hat{\alpha}_j}{\mu[j]^2}] \leq \max_j \frac{1}{\mu[j]^2} \left(\delta + \alpha_j + 2 \sqrt{\frac{\ln(d/\delta)}{m}} \cdot \max \left(\alpha_j, \frac{\ln(d/\delta)}{m} \right) \right)$$

Choose $\delta = 1/m$ and let j be a maximizer of the above. Consider two cases. If $\alpha_j < \ln(dm)/m$ then

$$\mathbb{E}_S[\max_j \frac{\hat{\alpha}_j}{\mu[j]^2}] \leq \max_j \frac{1}{\mu[j]^2} \cdot \frac{4 \ln(dm)}{m}.$$

Otherwise,

$$\mathbb{E}_S[\max_j \frac{\hat{\alpha}_j}{\mu[j]^2}] \leq \max_j \frac{1}{\mu[j]^2} (\delta + 3\alpha_j) \leq \max_j \frac{4\alpha_j}{\mu[j]^2}.$$

All in all, we have shown

$$R_m^L(U^\mu, D) \leq \max_j \frac{1}{\mu[j]} \sqrt{\frac{32 \ln(d)}{m}} \cdot \max\left(\alpha_j, \frac{\ln(dm)}{m}\right).$$

■

We now prove the desired Rademacher complexity bound on U_b .

Theorem 10 *For any distribution D over $(X, Y) \in [0, 1]^d$, if $dm \geq 3$,*

$$\mathcal{R}_m^L(U_b, D) \leq \sqrt{\frac{128k \ln(d)}{m}} \max\left(b, \frac{k \ln(dm)}{m}\right).$$

Proof Define α_j and U^μ as in Lemma 9. Let $J = \{j \in [d] \mid \alpha_j \geq \frac{b}{k}\}$, and $\bar{J} = \{j \in [d] \mid \alpha_j < \frac{b}{k}\}$. For a vector $v \in \mathbb{R}^d$ and a set $I \subseteq [d]$, denote by $v[I]$ the vector which is obtained from v by setting the coordinates not in I to zero. We have

$$\begin{aligned} \mathcal{R}_m^L(U_b, D) &= \frac{2}{m} \mathbb{E} \left[\sup_{w \in W} \left| \sum_{i=1}^m \epsilon_i Y_i \langle w, X_i \rangle \right| \right] \\ &= \frac{2}{m} \mathbb{E} \left[\sup_{w \in W} \left| \sum_{i=1}^m \epsilon_i Y_i \langle w[J], X_i[J] \rangle + \sum_{i=1}^m \epsilon_i Y_i \langle w[\bar{J}], X_i[\bar{J}] \rangle \right| \right] \\ &\leq \frac{2}{m} \mathbb{E} \left[\sup_{w \in W} \left| \sum_{i=1}^m \epsilon_i Y_i \langle w[J], X_i[J] \rangle \right| \right] + \frac{2}{m} \mathbb{E} \left[\sup_{w \in W} \left| \sum_{i=1}^m \epsilon_i Y_i \langle w[\bar{J}], X_i[\bar{J}] \rangle \right| \right] \\ &= \mathcal{R}_m^L(U_b, D_1) + \mathcal{R}_m^L(U_b, D_2), \end{aligned} \tag{17}$$

where D_1 is the distribution of $(X[J], Y)$ and D_2 is the distribution of $(X[\bar{J}], Y)$. We now bound the two Rademacher complexities of the right-hand side using Lemma 9.

To bound $\mathcal{R}_m^L(U_b, D_1)$, define $\mu_1 \in \mathbb{R}_+^d$ by $\mu_1[j] = \alpha_j/b$. It is easy to see that $U_b \subseteq U^{\mu_1}$. Therefore $\mathcal{R}_m^L(U_b, D_1) \leq \mathcal{R}_m^L(U^{\mu_1}, D_1)$. By Lemma 9 and the definition of μ_1

$$\begin{aligned} \mathcal{R}_m^L(U^{\mu_1}) &\leq \max_{j \in J} \frac{1}{\mu_1[j]} \sqrt{\frac{32 \ln(d)}{m}} \max\left(\alpha_j, \frac{\ln(dm)}{m}\right) \\ &= \max_{j \in J} \frac{b}{\alpha_j} \sqrt{\frac{32 \ln(d)}{m}} \max\left(\alpha_j, \frac{\ln(dm)}{m}\right) \\ &= \max_{j \in J} \sqrt{\frac{b}{\alpha_j} \frac{32 \ln(d)}{m}} \max\left(b, \frac{b}{\alpha_j} \frac{\ln(dm)}{m}\right). \end{aligned}$$

By the definition of J , for all $j \in J$ we have $\frac{b}{\alpha_j} \leq k$. It follows that

$$\mathcal{R}_m^L(U^{\mu_1}, D_1) \leq \sqrt{\frac{32k \ln(d)}{m}} \max\left(b, \frac{k \ln(dm)}{m}\right). \tag{18}$$

To bound $\mathcal{R}_m^L(U_b, D_2)$, define $\mu_2 \in \mathbb{R}_+^d$ by $\mu_2[j] = \frac{1}{k}$. Note that $U^{\mu_2} = W_k$ and $U_b \subseteq W_k$, hence $\mathcal{R}_m^L(U_b, D_2) \leq \mathcal{R}_m^L(U^{\mu_2}, D_2)$. By Lemma 9 and the definition of μ_2

$$\begin{aligned} \mathcal{R}_m^L(U^{\mu_2}, D_2) &\leq \max_{j \in \bar{J}} \frac{1}{\mu_2[j]} \sqrt{\frac{32 \ln(d)}{m} \max \left(\alpha_j, \frac{\ln(dm)}{m} \right)} \\ &= \max_{j \in \bar{J}} \sqrt{\frac{32k \ln(d)}{m} \max \left(k\alpha_j, \frac{k \ln(dm)}{m} \right)}. \end{aligned}$$

By the definition of \bar{J} , for all $j \in J$ we have $k\alpha_j \leq b$. Therefore

$$\mathcal{R}_m^L(U^{\mu_2}, D_2) \leq \sqrt{\frac{32k \ln(d)}{m} \max \left(b, \frac{k \ln(dm)}{m} \right)}. \quad (19)$$

Combining Eq. (17), Eq. (18) and Eq. (19) we get the statement of the theorem. \blacksquare

We can now derive our convergence result for negative examples.

Corollary 11 *Let $b \geq 0$. There exists a universal constant C such that for any distribution D over $[0, 1]^d \times \{\pm 1\}$ that draws only negative labels, with probability $1 - \delta$ over samples of size m , for any $w \in U_b$,*

$$\ell_-(w) \leq \hat{\ell}_-(w) + C \left(\sqrt{\frac{(kb + |r'|) \ln(edm/\delta)}{m}} + \frac{k \ln(edm/\delta)}{m} \right). \quad (20)$$

Proof Define $\phi : \mathbb{R} \rightarrow \mathbb{R}$ by $\phi(z) = [r' - z]_+$. Since D draws only negative labels, the Winnow loss on pairs (X, Y) drawn from D is exactly $\phi(Y \langle w, X \rangle)$. Note that ϕ is an application of a 1-Lipschitz function to a translation by r' of the linear loss. Thus, by the properties of the Rademacher complexity and by Theorem 10 we have, for $dm \geq 3$,

$$\begin{aligned} \mathcal{R}_m(U_b, D) &\leq \mathcal{R}_m^L(U_b, D) + \sqrt{\frac{|r'|}{m}} \\ &\leq \sqrt{\frac{128k \ln(d)}{m} \max \left(b, \frac{k \ln(dm)}{m} \right)} + \sqrt{\frac{|r'|}{m}}. \end{aligned} \quad (21)$$

Assume that $r' \leq 0$. By Talagrand's inequality (see e.g. Boucheron et al., 2005, Theorem 5.4), with probability $1 - \delta$ over samples of size m drawn from D , for all $w \in U_b$

$$\ell_-(w) \leq \hat{\ell}_-(w) + 2\mathcal{R}_m(U_b, D) + \sqrt{\frac{2 \sup_{w \in U_b} \text{Var}[\ell(X, Y, w)] \ln(1/\delta)}{m}} + \frac{4k \ln(1/\delta)}{3m}. \quad (22)$$

To bound the variance of $\ell(X, Y, w)$, we note that $\ell(X, Y, w) \in [0, k]$. In addition, $Y = -1$, thus $\ell(X, Y, w) = [r' + \langle w, X \rangle]_+$. Since $r' \leq 0$, for any $w \in U_b$

$$\text{Var}[\ell_-(X, w)] \leq k \cdot \mathbb{E}[\ell_-(X, w)] \leq k \cdot \mathbb{E}[\langle w, X \rangle] \leq kb. \quad (23)$$

Combining Eq. (21), Eq. (22) and Eq. (23) we conclude that there exists a universal constant C such that for any $w \in U_b$,

$$\ell_-(w) \leq \hat{\ell}_-(w) + C \left(\sqrt{\frac{(kb + |r'|) \ln(edm/\delta)}{m}} + \frac{k \ln(edm/\delta)}{m} \right).$$

Now, for any $r' > 0$, the values of $\hat{\ell}_-(w)$ and $\ell_-(w)$ are the same as the values for $r' = 0$ except for an identical additive term of r' , thus the same result holds. \blacksquare

4.2 Convergence on Positive Labels

For positive labels, we show a uniform convergence result. The idea of the proof technique below is as follows. First, following a technique in the spirit of the one given in Zhang (2002), we show that the regret bound for the online learning algorithm presented in Section 3 can be used to construct a small cover of the set of loss functions parameterized by W_k . Second, we convert the bound on the size of the cover to a bound on the Rademacher complexity, thus showing a uniform convergence result. This argument is a refinement of Dudley's entropy bound (Dudley, 1967), which is stated the most explicitly in Srebro et al. (2010) (Lemma A.3)

We start with the following direct corollary of Theorem 3:

Corollary 12 *Assume that the conditions of Theorem 3 hold. Assume also that there is \mathbf{u} such that $f_t(\mathbf{u}) = 0$ for all t . Set $\eta = \sqrt{\frac{2k \ln(kd)}{\beta T}}$ and assume that T is large enough so that $\alpha\eta \leq 1/2$. Then,*

$$\sum_{t=1}^T f_t(\mathbf{w}_t) \leq 4\sqrt{2\beta k \ln(kd)T}.$$

Let $k \geq r \geq 0$ be two real numbers and let $W \subseteq \mathbb{R}_+^d$. Let $f_{\mathbf{w}}$ denote the function defined by

$$f_{\mathbf{w}}(\mathbf{x}, y) = \ell(\mathbf{x}, y, \mathbf{w}),$$

and consider the class of functions

$$F_W = \{f_{\mathbf{w}} \mid \mathbf{w} \in W\}. \quad (24)$$

Given $S = ((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m))$, where $\mathbf{x}_i \in [0, 1]^d$ and $y_i \in \{\pm 1\}$, we say that (F_W, S) is (∞, ϵ) -properly-covered by a set $V \subseteq F_W$ if for any $f \in F_W$ there is a $g \in V$ such that

$$\|(f(\mathbf{x}_1, y_1), \dots, f(\mathbf{x}_m, y_m)) - (g(\mathbf{x}_1, y_1), \dots, g(\mathbf{x}_m, y_m))\|_{\infty} \leq \epsilon.$$

We denote by $\mathbb{N}_{\infty}(W, S, \epsilon)$ the minimum value of an integer N such that exists a $V \subseteq F_W$ of size N that (∞, ϵ) -properly-covers (F_W, S) .

The following lemma bounds the covering number for F_W , for sets S with all-positive labels y_i .

Lemma 13 *Let $S = ((\mathbf{x}_1, 1), \dots, (\mathbf{x}_m, 1))$, where $\mathbf{x}_i \in [0, 1]^d$, and let F_W be as defined in Eq. (24). Then,*

$$\ln \mathbb{N}_{\infty}(W_k, S, \epsilon) \leq C \cdot rk \ln(kd) \ln(m)/\epsilon^2.$$

Proof We use a technique in the spirit of the one given in Zhang (2002). Fix some \mathbf{u} , with $\mathbf{u} \geq 0$ and $\|\mathbf{u}\|_1 \leq k$. For each i let

$$g_i^{\mathbf{u}}(\mathbf{w}) = \begin{cases} |\langle \mathbf{w}, \mathbf{x}_i \rangle - \langle \mathbf{u}, \mathbf{x}_i \rangle| & \text{if } \langle \mathbf{u}, \mathbf{x}_i \rangle \leq r \\ [r - \langle \mathbf{w}, \mathbf{x}_i \rangle]_+ & \text{o.w.} \end{cases}$$

and define the function

$$G_{\mathbf{u}}(\mathbf{w}) = \max_i g_i^{\mathbf{u}}(\mathbf{w}).$$

It is easy to verify that for any \mathbf{w} ,

$$\|(f_{\mathbf{w}}(\mathbf{x}_1, 1), \dots, f_{\mathbf{w}}(\mathbf{x}_m, 1)) - (f_{\mathbf{u}}(\mathbf{x}_1, 1), \dots, f_{\mathbf{u}}(\mathbf{x}_m, 1))\|_{\infty} \leq G_{\mathbf{u}}(\mathbf{w}).$$

Now, clearly, $G_{\mathbf{u}}(\mathbf{u}) = 0$. In addition, for any $\mathbf{w} \geq 0$, a sub-gradient of $G_{\mathbf{u}}$ at \mathbf{w} is obtained by choosing i that maximizes $g_i^{\mathbf{u}}(\mathbf{w})$ and then taking a sub-gradient of $g_i^{\mathbf{u}}$, which is of the form $\mathbf{z} = \alpha \mathbf{x}_i$ where $\alpha \in \{-1, 0, 1\}$. If $\alpha \in \{-1, 1\}$, it is easy to verify that

$$\sum_j w[j] z[j]^2 \leq \langle \mathbf{w}, \mathbf{x}_i \rangle \leq g_i^{\mathbf{u}}(\mathbf{w}) + r = G_{\mathbf{u}}(\mathbf{w}) + r.$$

If $\alpha = 0$ then clearly $\sum_j w[j] z[j]^2 \leq G_{\mathbf{u}}(\mathbf{w}) + r$ as well.

We can now apply Cor. 12 by setting $f_t = G_{\mathbf{u}}$ for all t , setting $\alpha = 1$ and $\beta = r$ in Eq. (10), and noting that since $\mathbf{x}_i \in [0, 1]^d$, we have $\mathbf{z}_t \in [-1, 1]^d$ for all t . If $\eta \leq 1$ we have $\eta z_t[i] \geq -1$ for all t, i as needed. Since $\eta = \sqrt{\frac{2k \ln(kd)}{rT}}$, this holds for all $T \geq 2k \ln(kd)/r$.

We conclude that if we run the unnormalized EG algorithm with $T \geq 2k \ln(kd)/r$ and η and λ as required, we get

$$\sum_{t=1}^T G_{\mathbf{u}}(\mathbf{w}_t) \leq C \cdot \sqrt{rk \ln(kd)T}.$$

Dividing by T and using Jensen's inequality we conclude

$$G_{\mathbf{u}}\left(\frac{1}{T} \sum_t \mathbf{w}_t\right) \leq C \cdot \sqrt{\frac{rk \ln(kd)}{T}}.$$

Denote $\mathbf{w}_{\mathbf{u}} = \frac{1}{T} \sum_t \mathbf{w}_t$. Setting $\epsilon = C \cdot \sqrt{\frac{rk \ln(kd)}{T}}$, it follows that the following set is a (∞, ϵ) -proper-cover for (F_{W_k}, S) :

$$V = \{\mathbf{w}_{\mathbf{u}} \mid \mathbf{u} \in W_k\}.$$

Now, we only have left to bound the size of V . Consider again the unnormalized EG algorithm. Since $z_t = \alpha \mathbf{x}_i$ for some $\alpha \in \{-1, 0, +1\}$ and $i \in \{1, \dots, m\}$, at each round of the algorithm there are only two choices to be made: the value of i and the value of α . Therefore, the number of different vectors produced by running unnormalized EG for T iterations on $G_{\mathbf{u}}$ for different values of \mathbf{u} is at most $(3m)^T$. Thus $|V| \leq (3m)^T$. By our definition of ϵ ,

$$\ln |V| \leq T \ln(3m) \leq C \cdot rk \ln(kd) \ln(m) / \epsilon^2.$$

This concludes our proof. ■

Using this result we can bound from above the covering number defined using the Euclidean norm: We say that (F_W, S) is $(2, \epsilon)$ -properly-covered by a set $V \subseteq F_W$ if for any $f \in F_W$ there is a $g \in V$ such that

$$\frac{1}{\sqrt{m}} \|(f(\mathbf{x}_1, y_1), \dots, f(\mathbf{x}_m, y_m)) - (g(\mathbf{x}_1, y_1), \dots, g(\mathbf{x}_m, y_m))\|_2 \leq \epsilon.$$

We denote by $\mathbb{N}_2(W, S, \epsilon)$ the minimum value of an integer N such that exists a $V \subseteq F_W$ of size N that $(2, \epsilon)$ -properly-covers (F_W, S) . It is easy to see that for any two vectors $u, v \in \mathbb{R}^m$, $\frac{1}{\sqrt{m}}\|u - v\|_2 \leq \|u - v\|_\infty$. It follows that for any W and S , we have $\mathbb{N}_2(W, S, \epsilon) \leq \mathbb{N}_\infty(W, S, \epsilon)$.

The \mathbb{N}_2 covering number can be used to bound the Rademacher complexity of (F_W, S) using a refinement of Dudley's entropy bound (Dudley, 1967), which is stated the most explicitly in Srebro et al. (2010) (Lemma A.3). The lemma states that for any $\epsilon \geq 0$,

$$\mathcal{R}(W, S) \leq 4\epsilon + \frac{10}{\sqrt{m}} \int_\epsilon^B \sqrt{\ln \mathbb{N}_2(W, S, \gamma)} d\gamma,$$

where B is an upper bound on the possible values of $f \in F_W$ on members of S . For S with all-positive labels we clearly have $B \leq r$.

Combining this with Lemma 13, we get

$$\mathcal{R}(W_k, S) \leq C \cdot \left(\epsilon + \frac{1}{\sqrt{m}} \int_\epsilon^r \sqrt{rk \ln(kd) \ln(m)/\gamma} d\gamma \right) = C \cdot \left(\epsilon + \sqrt{\frac{rk \ln(kd) \ln(m)}{\sqrt{m}}} \ln(r/\epsilon) \right).$$

Setting $\epsilon = r/m$ we get

$$\mathcal{R}(W_k, S) \leq C \cdot \sqrt{\frac{rk \ln(ekd) \ln^3(m)}{m}}.$$

Thus, for any $k, d, m \geq 1$, and any distribution D over $[0, 1]^d \times \{\pm 1\}$ that draws only positive labels, we have

$$\mathcal{R}_m(W_k, D) \leq C \left(\sqrt{\frac{rk \ln(ekd) \ln^3(m)}{m}} \right).$$

By Rademacher sample complexity bounds Bartlett and Mendelson (2002), and since ℓ for positive labels is bounded by r , we can immediately conclude the following:

Theorem 14 *Let $k \geq r \geq 0$. For any distribution D over $[0, 1]^d \times \{\pm 1\}$ that draws only positive labels, with probability $1 - \delta$ over samples of size m , for any $w \in W_k$,*

$$\begin{aligned} \ell_+(w) &\leq \hat{\ell}_+(w) + C \cdot \left(\sqrt{\frac{rk \ln(ekd) \ln^3(m)}{m}} + \sqrt{\frac{r^2 \ln(1/\delta)}{m}} \right) \\ &\leq \hat{\ell}_+(w) + C \cdot \left(\sqrt{\frac{rk (\ln(ekd) \ln^3(m) + \ln(1/\delta))}{m}} \right). \end{aligned}$$

4.3 Combining negative and positive losses

We have shown separate convergence rate results for the loss on positive labels and for the loss on negative labels. In this section we combine these results to achieve a convergence result for the full Winnow loss. For this, we need to adapt the convergence results achieved above to take into account the fraction of positive and negative labels in the true distribution as well as in the sample. The following theorems accomplish this for the negative and the positive cases.

Theorem 15 *There exists a universal constant C such that for any distribution D over $[0, 1]^d \times \{\pm 1\}$, with probability $1 - \delta$ over samples of size m*

$$p_+ \ell_+(\hat{w}) \leq \hat{p}_+ \hat{\ell}_+(\hat{w}) + C \cdot \sqrt{\frac{rk(\ln(kd) \ln^3(m) + \ln(3/\delta))}{m}}.$$

Proof First, if $p_+ \leq \frac{16 \ln(1/\delta)}{m}$ then the theorem trivially holds. Therefore we assume that $p_+ \geq \frac{16 \ln(1/\delta)}{m}$. We have

$$p_+ \ell_+(\hat{w}) = \hat{p}_+ \hat{\ell}_+(\hat{w}) + (p_+ - \hat{p}_+) \hat{\ell}_+(\hat{w}) + p_+ (\ell_+(\hat{w}) - \hat{\ell}_+(\hat{w})). \quad (25)$$

To prove the theorem, we will bound the two rightmost terms. First, to bound $(p_+ - \hat{p}_+) \hat{\ell}_+(\hat{w})$, note that by definition of the loss function for positive labels we have that $\hat{\ell}_+(\hat{w}) \in [0, r]$. Therefore, Bernstein's inequality (Eq. (8)) implies that with probability $1 - \delta/3$

$$(p_+ - \hat{p}_+) \hat{\ell}_+(\hat{w}) \leq 2r \sqrt{\frac{\ln(3/\delta)}{m} \max(p_+, \frac{\ln(3/\delta)}{m})} \leq \sqrt{\frac{4r \ln(3/\delta)}{m}}. \quad (26)$$

Second, to bound $p_+ (\ell_+(\hat{w}) - \hat{\ell}_+(\hat{w}))$, we apply Theorem 14 to the conditional distribution induced by D on X given $Y = 1$, to get that with probability $1 - \delta/3$

$$p_+ (\ell_+(\hat{w}) - \hat{\ell}_+(\hat{w})) \leq p_+ \cdot C \cdot \sqrt{\frac{rk(\ln(ekd) \ln^3(m) + \ln(3/\delta))}{m \hat{p}_+}}.$$

Using our assumption on p_+ we obtain from Lemma 1 that with probability $1 - \delta/3$, $p_+/\hat{p}_+ \leq 2$. Therefore, $p_+/\sqrt{\hat{p}_+} \leq \sqrt{2p_+} \leq \sqrt{2}$. Thus, with probability $1 - 2\delta/3$,

$$p_+ (\ell_+(\hat{w}) - \hat{\ell}_+(\hat{w})) \leq C \cdot \sqrt{\frac{rk(\ln(ekd) \ln^3(m) + \ln(3/\delta))}{m}}. \quad (27)$$

Combining Eq. (25), Eq. (26) and Eq. (27) and applying the union bound, we get the theorem. \blacksquare

Theorem 16 *There exists a universal constant C such that for any distribution D over $[0, 1]^d \times \{\pm 1\}$, with probability $1 - \delta$ over samples of size m*

$$p_- \ell_-(\hat{w}) \leq \hat{p}_- \hat{\ell}_-(\hat{w}) + C \left(\sqrt{\frac{kr \ln(edm/\delta)}{m}} + \frac{k \ln(edm/\delta)}{m} \right). \quad (28)$$

Proof First, if $p_- \leq \frac{16 \ln(1/\delta)}{m}$ then the theorem trivially holds (since $\ell_-(\hat{w}) \in [0, r+k]$). Therefore we assume that $p_- \geq \frac{16 \ln(1/\delta)}{m}$. Thus, by Lemma 1, $\hat{p}_- \geq p_-/2$.

We have

$$p_- \ell_-(\hat{w}) = \hat{p}_- \hat{\ell}_-(\hat{w}) + (p_- - \hat{p}_-) \hat{\ell}_-(\hat{w}) + p_- (\ell_-(\hat{w}) - \hat{\ell}_-(\hat{w})). \quad (29)$$

To prove the theorem, we will bound the two rightmost terms. First, to bound $(p_- - \hat{p}_-) \hat{\ell}_-(\hat{w})$, note that by Bernstein's inequality and our assumption on p_- , with probability $1 - \delta$

$$p_- - \hat{p}_- \leq 2 \sqrt{\frac{\ln(1/\delta)}{m} \max(p_-, \frac{\ln(1/\delta)}{m})} = 2 \sqrt{\frac{p_- \ln(1/\delta)}{m}}.$$

By Lemma 6 and Lemma 1, $\hat{\ell}_-(\hat{w}) \leq \frac{2r}{\hat{p}_-} \leq \frac{4r}{p_-}$. In addition, by definition $\hat{\ell}_-(\hat{w}) \leq r+k \leq 2k$. Therefore

$$(p_- - \hat{p}_-) \hat{\ell}_-(\hat{w}) \leq 4 \min\left(\frac{2r}{p_-}, k\right) \sqrt{\frac{p_- \ln(1/\delta)}{m}}. \quad (30)$$

Now, if $k > 2r/p_-$, then the right-hand of the above becomes

$$8 \frac{r}{p_-} \sqrt{\frac{p_- \ln(1/\delta)}{m}} = 8 \sqrt{\frac{(r/p_-) \cdot r \ln(1/\delta)}{m}} \leq 8 \sqrt{\frac{k \cdot r \ln(1/\delta)}{m}}.$$

Otherwise, $k \leq 2r/p_-$ and the right-hand of Eq. (30) becomes

$$4k \sqrt{\frac{p_- \ln(1/\delta)}{m}} \leq 4k \sqrt{\frac{(2r/k) \ln(1/\delta)}{m}} \leq 8 \sqrt{\frac{k \cdot r \ln(1/\delta)}{m}}.$$

All in all, we have shown that

$$(p_- - \hat{p}_-) \hat{\ell}_-(\hat{w}) \leq 8 \sqrt{\frac{rk \ln(1/\delta)}{m}}. \quad (31)$$

Second, to bound $p_- (\ell_-(\hat{w}) - \hat{\ell}_-(\hat{w}))$, recall that by Theorem 8, we have

$$\hat{w} \in \{w \in \mathbb{R}_+^d \mid \|w\|_1 \leq k, \mathbb{E}_D[\langle w, X \rangle \mid Y = -1] \leq b\},$$

where b is defined as

$$b = \frac{8r}{p_-} + \frac{32k \ln(d/\delta)}{mp_-} \leq \frac{C}{p_-} (2r + \frac{k \ln(d/\delta)}{m}).$$

Thus, by Cor. 11, with probability $1 - \delta$

$$\ell_-(w) \leq \hat{\ell}_-(w) + C \left(\sqrt{\frac{(kb+r) \ln(edm\hat{p}_-/\delta)}{m\hat{p}_-}} + \frac{k \ln(edm\hat{p}_-/\delta)}{m\hat{p}_-} \right).$$

Since $\hat{p}_- \geq p_-/2$,

$$\ell_-(w) \leq \hat{\ell}_-(w) + C \left(\sqrt{\frac{(kb+r) \ln(edm/\delta)}{mp_-}} + \frac{k \ln(edm/\delta)}{mp_-} \right).$$

for some other constant C . Therefore, substituting b for its upper bound we get

$$p_-(\ell_-(w) - \hat{\ell}_-(w)) \leq C \left(\sqrt{\frac{kr \ln(edm/\delta)}{m}} + \frac{k \ln(edm/\delta)}{m} \right). \quad (32)$$

Combining Eq. (29), Eq. (31) and Eq. (32) we get the statement of the theorem. \blacksquare

Finally, we are ready to prove our main result for the sample complexity of ERM algorithms for Winnow.

Proof [of Theorem 5] From Theorem 15 and Theorem 16 we conclude that with probability $1 - \delta$,

$$\begin{aligned} \ell(\hat{w}) &= p_- \ell_-(\hat{w}) + p_+ \ell_+(\hat{w}) \\ &\leq \hat{p}_- \hat{\ell}_-(\hat{w}) + \hat{p}_+ \hat{\ell}_+(\hat{w}) + \sqrt{\frac{O(rk(\ln(kd) \ln^3(m) + \ln(1/\delta)))}{m}}. \end{aligned} \quad (33)$$

Now,

$$\hat{p}_- \hat{\ell}_-(\hat{w}) + \hat{p}_+ \hat{\ell}_+(\hat{w}) = \hat{\ell}(\hat{w}) \leq \hat{\ell}(w^*). \quad (34)$$

We have $\mathbb{E}[\ell(X, Y, w^*)] = \ell(w^*) \leq \ell(\mathbf{0}) \leq r$. Therefore, by Bernstein's inequality we have that with probability $1 - \delta$

$$\begin{aligned} \hat{\ell}(w^*) &= \hat{\mathbb{E}}[\ell(X, Y, w^*)] \leq \mathbb{E}[\ell(X, Y, w^*)] + \sqrt{\frac{\ln(1/\delta)}{m} \max\{\mathbb{E}[\ell(X, Y, w^*)], \frac{\ln(1/\delta)}{m}\}} \\ &\leq \ell(w^*) + \sqrt{\frac{r \ln(1/\delta)}{m}} + \frac{\ln(1/\delta)}{m}. \end{aligned}$$

Combining this with Eq. (34) we get that with probability $1 - \delta$

$$\hat{p}_- \hat{\ell}_-(\hat{w}) + \hat{p}_+ \hat{\ell}_+(\hat{w}) \leq \ell(w^*) + \sqrt{\frac{r \ln(1/\delta)}{m}} + \frac{\ln(1/\delta)}{m}.$$

In light of Eq. (33), we conclude Eq. (11) \blacksquare

5. Lower Bounds

In this section we provide lower bounds for the learning rate and for the uniform convergence rate of the Winnow loss ℓ_θ .

5.1 Learning rate lower bound

Fix a threshold θ . The best Winnow loss for a distribution D over $[0, 1]^d \times \{\pm 1\}$ using a hyperplane from a set $W \subseteq \mathbb{R}_+^d$ is denoted by $\ell_\theta^*(W) = \min_{w \in W} \ell_\theta(w)$. The following result shows that even if the data domain is restricted to the discrete domain $\{0, 1\}^d$, the number of samples required for learning with the Winnow loss grows at least linearly in θk .

Theorem 17 *Let $k \geq 1$ and let $\theta \in [1, k/2]$. The sample complexity of learning W_k with respect to the loss ℓ_θ is $\Omega(\theta k/\epsilon^2)$. That is, for all $\epsilon \in (0, 1/2)$ if the training set size is $m = o(\theta k/\epsilon^2)$, then for any learning algorithm, there exists a distribution such that the classifier, $h : \{0, 1\}^d \rightarrow \mathbb{R}_+$, that the algorithm outputs upon receiving m i.i.d. examples satisfies $\ell_\theta(h) - \ell_\theta^*(W_k) > \epsilon$ with a probability of at least $1/4$.*

In the following construction we use the notion of a *Hadamard matrix*. A Hadamard matrix of order n is an $n \times n$ matrix H_n with entries in $\{\pm 1\}$ such that $H_n H_n^T = nI_n$. In other words, such that all rows in the matrix are orthogonal to each other. Hadamard matrices exist at least for each n which is a power of 2 (Sylvester, 1867).

Lemma 18 *Assume k is a power of 2, and let $d = k^2$. Let $x_1, \dots, x_d \subseteq \{\pm 1\}^d$ be the rows of the Hadamard matrix of order d . For every $y \in \{\pm 1\}^d$, there exists a $w \in W' = \{w \in [-1, 1]^d \mid \|w\| \leq k\}$ such that for all $i \in [d]$, $y[i]\langle w, x_i \rangle = 1$.*

Proof By the definition of a Hadamard matrix, for all $i \neq j$, $\langle x_i, x_j \rangle = 0$. Given $y \in \{\pm 1\}^d$, set $w = \frac{1}{d} \sum_{j \in [d]} y_j x_j$. Then for each i ,

$$y_i \langle w, x_i \rangle = y_i \frac{1}{d} \sum_{j \in [d]} y_j \langle x_i, x_j \rangle = \frac{1}{d} y_i^2 \langle x_i, x_i \rangle = \frac{1}{d} \|x_i\|_2^2 = 1.$$

It is left to show that $w \in W'$. First, for all $i \in [d]$, we have

$$|w[i]| = \left| \frac{1}{d} \sum_{j \in [d]} y_j x_j[i] \right| \leq \frac{1}{d} \sum_{j \in [d]} |x_j[i]| = 1,$$

which yields $w \in [-1, 1]^d$. Second, using $\|w\|_1 \leq \sqrt{d}\|w\|_2$ and

$$\|w\|_2^2 = \langle w, w \rangle = \frac{1}{d^2} \sum_{i, j \in [d]} \langle y_i x_i, y_j x_j \rangle = \frac{1}{d^2} \sum_{i \in [d]} y_i^2 \langle x_i, x_i \rangle = \frac{1}{d^2} \sum_{i \in [d]} d = 1,$$

we obtain that $\|w\|_1 \leq \sqrt{d} = k$. ■

Lemma 19 *Let k be a power of 2 and let $d = 2k^2 + 1$. There is a set $\{x_1, \dots, x_{k^2}\} \subseteq \{0, 1\}^d$ such that for every $y \in \{\pm 1\}^{k^2}$, there exists a $w \in W_k$ such that for all $i \in [k^2]$, $y[i](\langle w, x_i \rangle - k/2) = \frac{1}{2}$.*

Proof From Lemma 18 we have that there is a set $X = \{x_1, \dots, x_{k^2}\} \subseteq \{\pm 1\}^{k^2}$ such that for each labeling $y \in \{\pm 1\}^{k^2}$, there exists a $w_y \in [-1, 1]^d$ with $\|w_y\|_1 \leq k$ such that for all $i \in [k^2]$, $y[i]\langle w_y, x_i \rangle = 1$. We now define a new set $\tilde{X} = \{\tilde{x}_1, \dots, \tilde{x}_{k^2}\} \subseteq \{0, 1\}^d$ based on X that satisfies the requirements of the lemma.

For each $i \in [k^2]$ let $\tilde{x}_i = [\frac{\vec{1} + x_i}{2}, \frac{\vec{1} - x_i}{2}, 1]$, where $[\cdot, \cdot, \cdot]$ denotes a concatenation of vectors and $\vec{1}$ is the all-ones vector. In words, each of the first k^2 coordinates in \tilde{x}_i is 1 if the corresponding coordinate in x_i is 1, and zero otherwise. Each of the next k^2 coordinates in \tilde{x}_i is 1 if the corresponding coordinate in x_i is -1 , and zero otherwise. The last coordinate in \tilde{x}_i is always 1.

Now, let $y \in \{\pm 1\}^{k^2}$ be a desired labeling. We defined \tilde{w}_y based on w_y as follows: $\tilde{w}_y = [[w_y]_+, [-w_y]_+, \frac{k - \|w_y\|_1}{2}]$, where by $z = [v]_+$ we mean that $z[j] = \max\{v[j], 0\}$. In words, the first k^2 coordinates of \tilde{w}_y are copies of the positive coordinates of w_y , with zero in the negative coordinates, and the next k^2 coordinates of \tilde{w}_y are the absolute values of the negative coordinates of w_y , with zero in the positive coordinates. The last coordinate is a scaling term.

We now show that \tilde{w}_y has the desired property on \tilde{X} . For each $i \in [k^2]$,

$$\begin{aligned} \langle \tilde{w}_y, \tilde{x}_i \rangle &= \langle \frac{\vec{1} + x_i}{2}, [w_y]_+ \rangle + \langle \frac{\vec{1} - x_i}{2}, [-w_y]_+ \rangle + \frac{k - |w_y|_1}{2} \\ &= |w_y|_1/2 + \langle x_i, w_y \rangle/2 + \frac{k - |w_y|_1}{2} = \langle x_i, w_y \rangle/2 + k/2 = y_i/2 + k/2. \end{aligned}$$

It follows that $y_i(\langle \tilde{w}_y, \tilde{x}_i \rangle - k/2) = y_i^2/2 = 1/2$.

Now, clearly $\tilde{w}_y \in \mathbb{R}_+^d$. In addition,

$$\|\tilde{w}_y\|_1 = \|w_y\|_1 + \frac{k - \|w_y\|_1}{2} = \|w_y\|_1/2 + k/2 \leq k.$$

Hence $\tilde{w}_y \in W_k$ as desired. ■

Lemma 20 *Let z be a power of 2 and let k such that z divides k . Let $d = 2kz + k/z$. There is a set $\{x_1, \dots, x_{zk}\} \subseteq \{0, 1\}^d$ such that for every $y \in \{\pm 1\}^{zk}$, there exists a $w \in W_k$ such that for all $i \in [zk]$, $y[i](\langle w, x_i \rangle - z/2) = \frac{1}{2}$.*

Proof By Lemma 19 there is a set $X = \{x_1, \dots, x_{z^2}\} \subseteq \{0, 1\}^{2z^2+1}$ such that for all $y \in \{\pm 1\}^{z^2}$, there exists a $w_y \in \mathbb{R}_+^{2z^2+1}$ such that $\|w_y\|_1 \leq z$ and for all $i \in [z^2]$, $y[i](\langle w_y, x_i \rangle - z/2) = \frac{1}{2}$.

We now construct a new set $\tilde{X} = \{\tilde{x}_1, \dots, \tilde{x}_{zk}\} \subseteq \{0, 1\}^{2kz+k/z}$ as follows: For $i \in [zk]$, let $n = \lfloor i/z^2 \rfloor$ and $m = i \bmod z^2$, so that $i = nz^2 + m$. The vector \tilde{x}_i is the concatenation of $\frac{kz}{z^2} = \frac{k}{z}$ vectors, each of which is of dimension $2z^2 + 1$, where all the vectors are the all-zeros vector, except the $(n+1)$ 'th vector which equals to x_{m+1} . That is:

$$\tilde{x}_i = [\underbrace{0}_{\in \mathbb{R}^{2z^2+1}}, \dots, \underbrace{0}_{\in \mathbb{R}^{2z^2+1}}, \underbrace{x_{m+1}}_{\text{block } n+1}, \underbrace{0}_{\in \mathbb{R}^{2z^2+1}}, \dots, \underbrace{0}_{\in \mathbb{R}^{2z^2+1}}] \in \mathbb{R}^{\frac{k}{z}(2z^2+1)}.$$

Given $\tilde{y} \in \{\pm 1\}^{kz}$, let us rewrite it as a concatenation of k/z vectors, each of which in $\{\pm 1\}^{z^2}$, namely,

$$\tilde{y} = [\underbrace{\tilde{y}(1)}_{\in \{\pm 1\}^{z^2}}, \dots, \underbrace{\tilde{y}(k/z)}_{\in \{\pm 1\}^{z^2}}] \in \{\pm 1\}^{kz}.$$

Define $\tilde{w}_{\tilde{y}}$ as the concatenation of k/z vectors in $\{\pm 1\}^{z^2}$, using w_y defined above for each $y \in \{\pm 1\}^{z^2}$, as follows:

$$\tilde{w}_{\tilde{y}} = [\underbrace{w_{\tilde{y}(1)}}_{\in \mathbb{R}_+^{2z^2+1}}, \dots, \underbrace{w_{\tilde{y}(k/z)}}_{\in \mathbb{R}_+^{2z^2+1}}] \in \mathbb{R}^{\frac{k}{z}(2z^2+1)}.$$

For each i such that $n = \lfloor i/z^2 \rfloor$ and $m = i \bmod z^2$, we have

$$\langle \tilde{w}_{\tilde{y}}, \tilde{x}_i \rangle - z/2 = \langle w_{\tilde{y}(n+1)}, x_{m+1} \rangle - z/2 = \frac{1}{2} \tilde{y}(n+1)[m+1].$$

Now $\tilde{y}(n+1)[m+1] = \tilde{y}[i]$, thus we get $\tilde{y}[i](\langle \tilde{w}_{\tilde{y}}, \tilde{x}_i \rangle - z/2) = \frac{1}{2}$ as desired. Finally, we observe that $\|\tilde{w}_{\tilde{y}}\|_1 = \sum_{n \in [k/z]} \|w_{\tilde{y}(n)}\|_1 \leq k/z \cdot z = k$, hence $\tilde{w}_{\tilde{y}} \in W_k$. \blacksquare

Proof [of Theorem 17] Let $k \geq 1$, $\theta \in [\frac{1}{2}, \frac{k}{2}]$. Define $z = 2\theta$. Let $n = \max\{n \mid 2^n \leq z\}$, and let $m = \max\{m \mid m2^n \leq k\}$. Define $\tilde{z} = 2^n$ and $\tilde{k} = m2^n$. We have that \tilde{z} is a power of 2 and \tilde{z} divides \tilde{k} . Let $\tilde{d} = 2\tilde{k}\tilde{z} + \tilde{k}/\tilde{z}$. By Lemma 20, there is a set $X = \{x_1, \dots, x_{\tilde{k}}\} \subseteq \{0, 1\}^{\tilde{d}}$ such that for every $y \in \{\pm 1\}^{|X|}$, there exists a $w_y \in W_k$ such that for all $i \in [\tilde{z}\tilde{k}]$, $y[i](\langle w_y, x_i \rangle - \tilde{z}/2) = \frac{1}{2}$.

Now, let $d = \tilde{d} + 1$, and define $\tilde{w}_y = [w_y, \frac{z-\tilde{z}}{2}]$ and $\tilde{x}_i = [x_i, 1]$. It follows that

$$\begin{aligned} y[i](\langle \tilde{w}_y, \tilde{x}_i \rangle - \theta) &= y[i](\langle \tilde{w}_y, \tilde{x}_i \rangle - z/2) \\ &= y[i](\langle w_y, x_i \rangle + z/2 - \tilde{z}/2 - z/2) \\ &= y[i](\langle w_y, x_i \rangle - \tilde{z}/2) = \frac{1}{2}. \end{aligned}$$

We conclude that for all $i \in [\tilde{z}\tilde{k}]$, $\ell_\theta(\tilde{x}_i, y[i], \tilde{w}_y) = 0$ and $\ell_\theta(\tilde{x}_i, 1 - y[i], \tilde{w}_y) = 1$. Moreover, $\text{sign}(\langle \tilde{w}_y, \tilde{x}_i \rangle - \theta) = y[i]$.

Now, for a given w define $h_w(x) = \text{sign}(\langle w, x_i \rangle - \theta)$, and consider the binary hypothesis class $H = \{h_w \mid w \in W_k\}$ over the domain X . Our construction of \tilde{w}_y shows that the set X is shattered by this hypothesis class, thus its VC dimension is at least $|X|$. By VC-dimension lower bounds (e.g. Anthony and Bartlett 1999, Theorem 5.2), it follows that for any learning algorithm for H , if the training set size is $o(|X|/\epsilon^2)$, then there exists a distribution over X so that with probability greater than $1/64$, the output \hat{h} of the algorithm satisfies

$$\mathbb{E}[\hat{h}(x) \neq y] > \min_{w \in W_k} \mathbb{E}[h_w(x) \neq y] + \epsilon. \quad (35)$$

Next, we show that the existence of a learning algorithm for W_k with respect to ℓ_θ whose sample complexity is $o(|X|/\epsilon^2)$ would contradict the above statement. Indeed, let w^* be a minimizer of the right-hand side of Eq. (35), and let y^* be the vector of predictions of w^* on X . As our construction of \tilde{w}_{y^*} shows, we have $\ell_\theta(\tilde{w}_{y^*}) = \mathbb{E}[h_{w^*}(x) \neq y]$. Now, suppose that some algorithm learns $\hat{w} \in W_k$ so that $\ell_\theta(\hat{w}) \leq \ell_\theta^*(W_k) + \epsilon$. This implies that

$$\ell_\theta(\hat{w}) \leq \ell_\theta(\tilde{w}_{y^*}) + \epsilon = \mathbb{E}[h_{w^*}(x) \neq y] + \epsilon.$$

In addition, define a (probabilistic) classifier, \hat{h} , that outputs the label $+1$ with probability $p(\hat{w}, x)$ where $p(\hat{w}, x) = \min\{1, \max\{0, 1/2 + (\langle \hat{w}, x \rangle - \theta)\}\}$. Then, it is easy to verify that

$$\mathbb{P}[\hat{h}(x) \neq y] \leq \ell_\theta(x, y, \hat{w}).$$

Therefore, $\mathbb{E}[\hat{h}(x) \neq y] \leq \ell_\theta(\hat{w})$, and we obtain that

$$\mathbb{E}[\hat{h}(x) \neq y] \leq \mathbb{E}[h_{w^*}(x) \neq y] + \epsilon,$$

which leads to the desired contradiction. \blacksquare

We next show that the uniform convergence rate for our problem is in fact slower than the achievable learning rate.

5.2 Uniform convergence lower bound

The next theorem shows that the rate of uniform convergence for our problem is too slow, even if the distribution draws only negative labels.

Theorem 21 *Let $k \geq 1$, and assume $\theta \leq k/2$. There exists a distribution D over $\{0, 1\}^{k^2+1} \times Y$ such that $\forall x \in \{0, 1\}^d, \mathbb{P}[Y = -1 \mid X = x] = 1$, and $\ell^*(W_k, D) = [r']_+$, and such that with probability at least $1/2$ over samples $S \sim D^m$,*

$$\exists w \in W_k, \quad |\ell(w, S) - \ell(w, D)| \geq \Omega(\sqrt{k^2/m}). \quad (36)$$

To prove this theorem we first show two useful lemmas. The first lemma shows that a lower bound for the Rademacher complexity of a function class implies a lower bound on the uniform convergence of this function class. The derivation is similar to the proof of the upper bound in Bartlett and Mendelson (2002).

Lemma 22 *Let Z be a set, and consider a function class $F \subseteq [0, 1]^Z$. Let D be a distribution over Z . If $\mathcal{R}_m(F, D) \geq \alpha$, then with probability at least $1 - \delta$ over samples $S \sim D^m$,*

$$\exists f \in F, \quad |\mathbb{E}_{X \sim S}[f(X)] - \mathbb{E}_{X \sim D}[f(X)]| \geq \alpha/2 - \sqrt{\frac{\ln(1/\delta)}{8m}}.$$

Proof Denote $E[f, S] = \mathbb{E}_{X \sim S}[f(X)]$, and $E[f, D] = \mathbb{E}_{X \sim D}[f(X)]$. Consider two independent samples $S = (X_1, \dots, X_m), S' = (X'_1, \dots, X'_m) \sim D^m$, and let $\sigma = (\sigma_1, \dots, \sigma_m)$ be m independent random variables drawn uniformly from $\{\pm 1\}$. We have

$$\begin{aligned} 2\mathbb{E}_S[\sup_{f \in F} |E[f, S] - E[f, D]|] &\geq \mathbb{E}_{S, S'}[\sup_{f \in F} |E[f, S] - E[f, D]| + \sup_{f \in F} |E[f, S'] - E[f, D]|] \\ &\geq \mathbb{E}_{S, S'}[\sup_{f \in F} |E[f, S] - E[f, D]| + |E[f, S'] - E[f, D]|] \\ &\geq \mathbb{E}_{S, S'}[\sup_{f \in F} |E[f, S] - E[f, S']|] \\ &= \frac{1}{m} \mathbb{E}_{S, S'}[\sup_{f \in F} |\sum_{i \in [m]} f(X_i) - f(X'_i)|] \\ &= \frac{1}{m} \mathbb{E}_{S, S'}[\sup_{f \in F} |\sum_{i \in [m]} f(X_i) - f(X'_i)|] \\ &= \frac{2}{m} \mathbb{E}_{\sigma, S}[\sup_{f \in F} |\sum_{i \in [m]} \sigma_i f(X_i)|] = \mathcal{R}_m(F, D). \end{aligned}$$

Thus by the assumed lower bound on the Rademacher complexity,

$$\mathbb{E}_S[\sup_{f \in F} |E[f, S] - E[f, D]|] \geq \alpha/2.$$

We have left to show a lower bound with high probability. Define $g(S) = \sup_{f \in F} |E[f, S] - E[f, D]|$. Any change of one element in S can cause $g(S)$ to change by at most $1/m$. Therefore, by McDiarmid's inequality, $\mathbb{P}[g(S) \leq \mathbb{E}[g(S)] - t] \leq \exp(-2mt^2)$. It follows that with probability at least $1 - \delta$,

$$\sup_{f \in F} |E[f, S] - E[f, D]| \geq \alpha/2 - \sqrt{\frac{\ln(1/\delta)}{8m}}.$$

■

The next lemma provides a uniform convergence lower bound for a universal class of binary functions.

Lemma 23 *There exist universal constants c, C, C' such that the following holds. Let $H = \{0, 1\}^{[n]}$ be the set of all binary functions on $[n]$. Let D be the uniform distribution $[n]$. For any $n \geq C'$, with probability of at least $\frac{1}{2}$ over i.i.d. samples of size m drawn from D ,*

$$\exists h \in H, \quad |\mathbb{E}_{X \sim S}[h(X)] - \mathbb{E}_{X \sim D}[h(X)]| \geq \max\{C \cdot \sqrt{\frac{n}{m}}, c\}.$$

Proof Denote $E[h, S] = \mathbb{E}_{X \sim S}[h(X)]$, and $E[h, D] = \mathbb{E}_{X \sim D}[h(X)]$.

First, consider the case $m/n < 8$. For a given sample S define $h_S \in \{\pm 1\}^n$ such that

$$h_S(j) = \mathbb{I}[j \text{ appears in } S],$$

and denote by $N(S)$ the number of elements from $[n]$ that do not appear in S . Then

$$E[h_S, D] = \frac{1}{n} \sum_{j \in [n]} h_S(j) = \frac{1}{n} \sum_{j \in [n]} \mathbb{I}[j \text{ appears in } S] = 1 - \frac{N(S)}{n}.$$

On the other hand, $E[h_S, S] = 1$. It follows that

$$|E[h_S, S] - E[h_S, D]| \geq N(S)/n.$$

Using the fact that $1 - x \geq \exp(-2x)$ for $x \leq 1/2$, we get that for $n > 1$,

$$\mathbb{E}_S[N(S)] = \sum_{j \in [n]} \mathbb{P}[j \text{ does not appear in } S] = n(1 - \frac{1}{n})^m \geq n \exp(-2m/n) \geq n \exp(-16).$$

It follows that $\mathbb{E}_S[E[h_S, S] - E[h_S, D]] \geq \exp(-16)$.

To show that this difference is high with high probability over the choice of S , denote $f(S) = E[h_S, S] - E[h_S, D]$. Any change of one element in S can cause $f(S)$ to change by at most $1/n$. Therefore, by McDiarmid's inequality, $\mathbb{P}[f(S) \leq \mathbb{E}[f(S)] - t] \leq \exp(-2n^2 t^2 / m)$. It follows that with probability at least $1/2$,

$$f(S) \geq \exp(-16) - \sqrt{\frac{\ln(2)m}{2n^2}} \geq \exp(-16) - \sqrt{\frac{4 \ln(2)}{n}},$$

where the last inequality follows from the assumption that $m/n < 8$. It follows that there are constants $c, C > 0$ such that $n > C$, with probability of at least $1/2$, $E[h_S, S] - E[h_S, D] \geq c$.

Second, consider the case $m/n \geq 8$. By Lemma 22, it suffices to provide a lower bound for $\mathcal{R}_m(H, D)$. Fix a sample $S = (x_1, \dots, x_m)$ drawn from D . We have

$$m\mathcal{R}(H, S) = \mathbb{E}_\sigma \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i h(x_i) \right].$$

For a given $\sigma \in \{\pm 1\}^m$, define $h_\sigma \in H$ such that $h_\sigma(j) = \text{sign}(\sum_{i:x_i=j} \sigma_i)$. Then

$$\begin{aligned} \frac{m}{2} \mathcal{R}(F, S) &\geq \mathbb{E}_\sigma \left[\left| \sum_{i \in [m]} \sigma_i h_\sigma(x_i) \right| \right] \\ &= \mathbb{E}_\sigma \left[\left| \sum_{j \in [n]} \sum_{i:x_i=j} \sigma_i h_\sigma(j) \right| \right] \\ &= \mathbb{E}_\sigma \left[\left| \sum_{j \in [n]} \sum_{i:x_i=j} \sigma_i \text{sign} \left(\sum_{i:x_i=j} \sigma_i \right) \right| \right] \\ &= \sum_{j \in [n]} \mathbb{E}_\sigma \left[\left| \sum_{i:x_i=j} \sigma_i \right| \right]. \end{aligned}$$

Now, let $c_j(S) = |i : x_i = j|$. The expression $\mathbb{E}_\sigma \left[\left| \sum_{i:x_i=j} \sigma_i \right| \right]$ is equal to the expected distance of a random walk of length $c_j(S)$, which can be bounded from below by $\sqrt{c_j(S)/2}$ (Szarek, 1976). Therefore,

$$\mathcal{R}(H, S) \geq \frac{1}{\sqrt{2}m} \sum_{j \in [k^2]} \sqrt{c_j(S)}.$$

Taking expectation over samples S drawn from D , we get

$$\mathcal{R}(H, D) = \mathbb{E}_{S \sim D^m} [\mathcal{R}(H, S)] \geq \frac{1}{\sqrt{2}m} \sum_{j \in [n]} \mathbb{E}_S \left[\sqrt{c_j(S)} \right]. \quad (37)$$

Our final step is to bound $\mathbb{E}_S \left[\sqrt{c_j(S)} \right]$. We have $\mathbb{E}_S[c_j(S)] = \frac{m}{n}$, and $\text{Var}_S[c_j(S)] = \frac{m}{n}(1 - \frac{1}{n})$. Thus, by Chebyshev's inequality,

$$\mathbb{P}[c_j(S) \leq \frac{m}{n} - t] \leq \frac{m(1 - 1/n)}{nt^2} \leq \frac{m}{nt^2}.$$

Therefore

$$\mathbb{E}_S \left[\sqrt{c_j(S)} \right] \geq \left(1 - \frac{m}{nt^2}\right) \sqrt{\frac{m}{n}} - t.$$

Setting $t = \sqrt{2m/n}$, we get

$$\mathbb{E}_S \left[\sqrt{c_j(S)} \right] \geq \frac{1}{2} \sqrt{\frac{m}{n}} - \sqrt{\frac{2m}{n}}.$$

Now, since $m/n \geq 8$ it is easy to check that $\mathbb{E}_S \left[\sqrt{c_j(S)} \right] \geq \sqrt{m/8n}$. Plugging this into Eq. (37), we get

$$\mathcal{R}(H, D) \geq \frac{1}{\sqrt{2}m} \sum_{j \in [n]} \sqrt{m/8n} = \frac{1}{4} \sqrt{\frac{n}{m}}.$$

By Lemma 22, it follows that with probability at least $1 - \delta$ over samples,

$$\exists f \in F, \quad |\mathbb{E}_{X \sim S}[f(X)] - \mathbb{E}_{X \sim D}[f(X)]| \geq \frac{1}{8} \sqrt{\frac{n}{m}} - \sqrt{\frac{\ln(1/\delta)}{8m}} = \frac{n/8 - \ln(1/\delta)}{8m}.$$

Fixing $\delta = 1/2$, we get the desired lower bound. ■

Using the two lemmas above, we are now ready to prove our uniform convergence lower bound. We do this by mapping a subset of W_k to a universal class of binary functions over $\Theta(k^2)$ elements from our domain. Note that for this lower bound it suffices to consider the more restricted domain of binary vectors.

Proof [of Theorem 21] Let q be the largest power of 2 such that $q \leq k$. By Lemma 19, there exists a set of vectors $Z = \{z_1, \dots, z_{q^2}\} \subseteq \{0, 1\}^{q^2+1}$ such that for every $t \in \{\pm 1\}^{q^2}$ there exists a $w_t \in W_k$ such that for all i , $t[i](\langle w, z_i \rangle - q/2) = \frac{1}{2}$. Denote $U = \{w_t \mid t \in \{\pm 1\}^{q^2}\}$. It suffices to prove a lower bound on the uniform convergence of U , since this implies the same lower bound for W_k . Define the distribution D over $Z \times \{\pm 1\}$ such that for $(X, Y) \sim D$, X is drawn uniformly from z_1, \dots, z_{q^2} and $Y = -1$ with probability 1.

Consider the set of functions $H = \{0, 1\}^Z$, and for $h \in H$ define $t_h \in \{\pm 1\}^{q^2}$ such that for all $i \in [q^2]$, $t_h[i] = 2h(z_i) - 1$. For any $i \in [q^2]$, we have

$$\ell(z_i, -1, w_{t_h}) = [r' + \langle w, z_i \rangle]_+ = [r' + (t[i] + k)/2]_+ = [r' + (k-1)/2 + h(i)]_+ = r' + (k-1)/2 + h(z_i).$$

The last equality follows since $r' \geq \frac{1-k}{2}$. It follows that for any $h \in H$ and any sample S drawn from D ,

$$|\ell(w_{t_h}, S) - \ell(w_{t_h}, D)| = |\mathbb{E}_{X \sim S}[h(X)] - \mathbb{E}_{X \sim D}[h(X)]|.$$

By Lemma 23, with probability of at least $\frac{1}{2}$ over the sample $S \sim D^m$,

$$\exists h \in H, \quad |\mathbb{E}_{X \sim S}[h(X)] - \mathbb{E}_{X \sim D}[h(X)]| \geq \Omega(\sqrt{q^2/m}) = \Omega(\sqrt{k^2/m}).$$

Thus, with probability at least $1/2$,

$$\exists w \in W_k, \quad |\ell(w_{t_h}, S) - \ell(w_{t_h}, D)| \geq \Omega(\sqrt{k^2/m}).$$
■

References

- N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. In *Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on*, pages 292–301. IEEE, 1993.
- M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
- P. Auer and M.K. Warmuth. Tracking the best disjunction. *Machine Learning*, 32(2):127–150, 1998.
- P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

- S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.
- R.M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290 – 330, 1967.
- Claudio Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53:265–299, 2003.
- S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Proceedings of NIPS*, 2009.
- J. Kivinen and M. Warmuth. Additive versus exponentiated gradient updates for learning linear functions. Technical Report UCSC-CRL-94-16, University of California Santa Cruz, Computer Research Laboratory, 1994.
- N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.
- R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Machine Learning: Proceedings of the Fourteenth International Conference*, pages 322–330, 1997. To appear, *The Annals of Statistics*.
- S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University, 2007.
- Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.
- N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low-noise and fast rates. *CoRR*, abs/1009.3896, 2010.
- N. Srebro, K. Sridharan, and A. Tewari. On the universality of online mirror descent. *Advances in Neural Information Processing Systems (NIPS)*, 2011.
- J.J. Sylvester. Thoughts on inverse orthogonal matrices, simultaneous signsuccessions, and tessellated pavements in two or more colours, with applications to newton’s rule, ornamental tile-work, and the theory of numbers. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 34(232):461–475, 1867.
- S.J. Szarek. On the best constants in the khinchin inequality. *Studia Math*, 58(2), 1976.
- T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.